ELSEVIER

Discrete Applied Mathematics 85 (1998) 37–61

DISCRETE APPLIED MATHEMATICS

Scale-sensitive dimensions and skeleton estimates for classification

Márta Horváth [a,*,1], Gábor Lugosi [b]

[a] *Department of Mathematics and Computer Science, Technical University of Budapest, 1521 Stoczek u. 2, Budapest, Hungary*

[b] *Department of Economics, Pompeu Fabra University, Ramon Trias Fargas, 25-27, 08005 Barcelona, Spain*

Received 18 Decmber 1996; received in revised form 16 May 1997; accepted 23 October 1997

Abstract

The classical binary classification problem is investigated when it is known in advance that the posterior probability function (or regression function) belongs to some class of functions. We introduce and analyze methods which effectively exploit this knowledge. These methods are based on minimizing the empirical risk over a carefully selected "skeleton" of the class of regression functions. The skeletons are coverings of the class based on metrics, especially fitted for classification. A new scale-sensitive dimension is introduced which is more suitable for the studied classification problem than other, previously defined, dimension measures. This fact is demonstrated by performance bounds for the skeleton estimates in terms of the new dimension.[2]

1. Definitions

The following pattern classification problem is investigated: let (X, Y) be a pair of random variables, taking their values from some set $\mathscr{X}$ and $\{0,1\}$, respectively. The value of the *label* Y is to be predicted upon observing the *feature vector* X. The *prediction rule* or *classifier* g is a function $\mathscr{X} \to \{0,1\}$, whose performance is measured by the *probability of error*

$$L(g) = P\{g(X) \neq Y\}.$$

The joint distribution of (X, Y) is determined by the *regression function*

$$\eta^*(x) = P\{Y = 1 | X = x\}$$

* Corresponding author. E-mail: marti@inf.bme.hu.

[1] The research of the author was supported by MTA SZTAKI.

[2] Parts of the paper were presented at COLT'96 [21].

0166-218X/98/$19.00
PII S0166-218X(98)00013-4

(also known as the *a posteriori probability function*) and the measure μ of X on $\mathscr{X}$, that is,

$$\mu(A) = P\{X \in A\} \quad \text{for each measurable set } A \subseteq \mathscr{X}.$$

The Bayes classifier

$$g^*(x) = \begin{cases} 0 & \text{if } \eta^*(x) < \frac{1}{2}, \\ 1 & \text{otherwise,} \end{cases}$$

is well-known to have minimal probability of error among all possible classifiers. Its error probability $L(g^*)$ is called the Bayes risk, and is denoted by L^*.

Recall that if $\eta : \mathscr{X} \to [0,1]$ is an arbitrary measurable function, and we define the corresponding classifier by

$$g(x) = \begin{cases} 0 & \text{if } \eta(x) < \frac{1}{2}, \\ 1 & \text{otherwise,} \end{cases}$$

then the following elementary property holds:

$$\begin{aligned} L(g) - L^* &= 2E\left\{I_{\{g(X) \neq g^*(X)\}} \left|\eta^*(X) - \tfrac{1}{2}\right|\right\} \\ &\leqslant 2E\left\{I_{\{g(X) \neq g^*(X)\}} \left|\eta^*(X) - \eta(X)\right|\right\}, \end{aligned} \tag{1}$$

see, for example, [13, p. 16]. (I_A denotes the indicator of an event A.)

Assume that n independent copies of (X, Y) form the available data sequence:

$$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n)).$$

These data may be used to obtain the classification rule $g_n(x)$, whose probability of error is the random variable

$$L(g_n) = P\{g_n(X) \neq Y | D_n\}.$$

Very often, apart from the training sequence, some prior information is available about the joint distribution of (X, Y). For example, in some applications with $\mathscr{X} = \mathscr{R}^d$, it is known that η^* is a monotone function in all components of x. In other situations it may be known that η^* is a smooth function. In the basic PAC-learning setup [10], η^* is known to be the indicator function of one of the sets in a given class of sets. We assume throughout that η^* is a member of a known class of functions $\mathscr{F}$. In this paper we are interested in how this extra information can be exploited to obtain small probabilities of error.

At this point we need to introduce some notation:

- If $\eta \in \mathscr{F}$ is a regression function, $L(\eta)$ denotes the probability of error $L(g)$ of the corresponding classifier $g(x) = I_{\{\eta(x) \geqslant 1/2\}}$. We will always denote $g'(x) = I_{\{\eta'(x) \geqslant 1/2\}}$, $g^*(x) = I_{\{\eta^*(x) \geqslant 1/2\}}$, $g_n(x) = I_{\{\eta_n(x) \geqslant 1/2\}}$, etc.
- V denotes the VC *dimension* of the class of classifiers induced by $\mathscr{F}$, that is, the VC dimension of the class of sets of the form $\{x : \eta(x) \geqslant \frac{1}{2}\}$, $\eta \in \mathscr{F}$.

- $S_{\mathscr{F}}(x_1^n)$ is the *shatter coefficient* of the class $\mathscr{F}$ restricted to $x_1^n=(x_1,\dots,x_n)$, that is, the number of different ways the members of $\mathscr{F}$ can classify the n points $x_1,\dots,x_n$.
- $d_{H,n}(\eta,\eta')=(1/n)\sum_{i=1}^n I_{\{g(x_i)\neq g'(x_i)\}}$ denotes the *empirical (normalized) Hamming distance* between two classifiers, and $d_{1,n}(\eta,\eta')=(1/n)\sum_{i=1}^n |\eta(X_i)-\eta'(X_i)|$ is the *empirical L_1 distance*.
- $d_H(\eta,\eta')=P\{g(X)\neq g'(X)\}$, and $d_1(\eta,\eta')=E\{|\eta(X)-\eta'(X)|\}$, denote the corresponding *theoretical distances*.
- $N_H(\varepsilon,x_1^n,\mathscr{F})$ is the *empirical Hamming covering number* of $\mathscr{F}$ restricted to the set $\{x_1,\dots,x_n\}$, that is, the set of functions of smallest cardinality satisfying the property that for every $\eta\in\mathscr{F}$ there is an η' in the set such that $d_{H,n}(\eta,\eta')<\varepsilon$. Note that for $\varepsilon>1/n$, $N_H(\varepsilon,x_1^n,\mathscr{F})\leqslant S_{\mathscr{F}}(x_1^n)$.
- The *empirical L_1 covering number* $N_1(\varepsilon,x_1^n,\mathscr{F})$ is defined similarly, replacing the distance $d_{H,n}$ above by the empirical L_1 distance $d_{1,n}$.
- The covering numbers of $\mathscr{F}$ with respect to the theoretical distances d_H and d_1 are denoted by $N_H(\varepsilon,\mu,\mathscr{F})$ and $N_1(\varepsilon,\mu,\mathscr{F})$, respectively.
- The *scale-sensitive dimension* P_γ of Kearns and Shapire [17] (also known as the *fat shattering function*) for $0<\gamma\leqslant\frac{1}{2}$ is defined as follows: we say that $\mathscr{F}$ *γ-fat-shatters* a finite set $A\subset\mathscr{X}$ if there exists some function $s: A\to[0,1]$ such that for every subset $E\subset A$ there is a function $\eta_E\in\mathscr{F}$ such that $\eta_E(x)\geqslant s(x)+\gamma$ if $x\in E$ and $\eta_E(x)\leqslant s(x)-\gamma$ if $x\in A-E$. P_γ of $\mathscr{F}$ is the largest positive integer n for which there exists a set A of cardinality n which is γ-fat-shattered by $\mathscr{F}$. If for every n there is a set A which is γ-fat-shattered by $\mathscr{F}$ then we say that $P_\gamma=\infty$.

2. Methods of empirical risk minimization

Perhaps the most natural way of exploiting the knowledge that η^* is in a known class of functions is the following: form the class of classifiers determined by the functions in $\mathscr{F}$, that is,

$$\{g: g(x)=I_{\{\eta(x)\geqslant 1/2\}}, \eta\in\mathscr{F}\}.$$

The Bayes classifier is clearly in this class. Then we may select a member of the class by minimizing the *empirical error*

$$L_n(g)=\frac{1}{n}\sum_{i=1}^n I_{\{g(X_i)\neq Y_i\}}.$$

Then an inequality of Vapnik and Chervonenkis [29] (see also [3]) implies that if $\tilde{g}_n$ is the selected classifier, then

$$E\{L(\tilde{g}_n)\}-L^*\leqslant K\max\left(\sqrt{\frac{L^*\log(E\{S_{\mathscr{F}}(X_1^n)\})}{n}},\frac{\log(E\{S_{\mathscr{F}}(X_1^n)\})}{n}\right). \tag{2}$$

Here, and in the rest of the paper, K denotes a universal constant. It is well-known [25] that the shatter coefficients may be uniformly bounded as

$$\sup_{x_1,\dots,x_n} S_{\mathcal{F}}(x_1^n) \leqslant \left(\frac{ne}{V}\right)^V .$$

Therefore, we have

$$E\{L(\tilde{g}_n)\} - L^* \leqslant K \max\left(\sqrt{\frac{L^* V \log n}{n}}, \frac{V \log n}{n}\right).$$

The method of minimizing the empirical error is extremely robust in the sense that the above upper bound remains true in a completely distribution-free setting if we replace L^* by the probability of error $\underline{L} = \inf_{\eta \in \mathcal{F}} L(\eta)$ of the best classifier in the class. In other words, if $\eta^* \notin \mathcal{F}$, the method still works reasonably well.

On the other hand, this result is optimal (up to a logarithmic factor) in a minimax sense: for *any* classification rule g_n, there exists a distribution of (X, Y) such that

$$EL(g_n) - \underline{L} \geqslant K \max\left(\sqrt{\frac{\underline{L}V}{n}}, \frac{V}{n}\right), \tag{3}$$

see [13, 14]. However, if we exploit the additional information that $\eta^* \in \mathcal{F}$, we may be able to define classifiers with improved performance guarantees. It is precisely the goal of this paper to explore this direction.

For simplicity, we have given bounds for the expected value of the probability of error $L(\tilde{g}_n)$. Alternatively, we may rephrase these results in terms of sample-size bounds. For example, another consequence of the above-mentioned result of Vapnik and Chervonenkis is that for any $\varepsilon, \delta > 0$, $P\{L(\tilde{g}_n) - L^* > \varepsilon\} < \delta$ if the sample size n is at least

$$K \max\left(\frac{L^* V}{\varepsilon^2} \log \frac{V}{\varepsilon} + \frac{L^*}{\varepsilon^2} \log \frac{1}{\delta}, \frac{V}{\varepsilon} \log \frac{V}{\varepsilon} + \frac{1}{\varepsilon} \log \frac{1}{\delta}\right).$$

Note that unless $L^* < \varepsilon$, the maximum is achieved by the first term.

The disadvantage of this method of empirical error minimization is that it only uses a part of the available information. Apart from the form of the classifiers defined by the functions in $\mathcal{F}$, no other information about the shape of these functions is exploited. For example, if $\mathcal{F}$ is the class of Lipschitz functions $\eta : [0,1] \to [0,1]$ (with Lipschitz constant 1), then clearly the class of classifiers defined by $\mathcal{F}$ is too large, and we have $V = \infty$.

Another approach to the classification problem is to directly estimate the regression function η^* by a function η_n'. For the probability of error of the corresponding classifier $g_n'(x) = I_{\{\eta_n'(x) \geqslant 1/2\}}$, we have from (1) and the Cauchy–Schwarz inequality that

$$EL(g_n') - L^* \leqslant 2E\{|\eta^*(X) - \eta_n'(X)|\} \leqslant 2\sqrt{E\{|\eta^*(X) - \eta_n'(X)|^2\}}. \tag{4}$$

Observing that $\eta^*(x)=E\{Y|X=x\}$ and therefore for any η

$$E\{(\eta(X)-\eta^*(X))^2\}=E\{(\eta(X)-Y)^2\}-E\{(\eta^*(X)-Y)^2\},$$

leads to the following estimate: minimize the empirical L_2 error

$$\frac{1}{n}\sum_{i=1}^{n}(\eta(X_i)-Y_i)^2 \quad \eta\in\mathscr{F},$$

and consider the classifier g_n' corresponding to the minimizing function η_n'. Then, exploiting that the regression function η^* is in $\mathscr{F}$, we may use a result of Lee et al. [19, Theorem 7] to conclude that for every $\varepsilon>0$,

$$EL(g_n')-L^*\leqslant K\left(\sup_{x_1,\ldots,x_n}\sqrt{\frac{\log N(\varepsilon,x_1^n,\mathscr{F})}{n}}+\varepsilon\right). \tag{5}$$

In Section 5 we present an alternative method with improved performance guarantees.

As proved by Alon et al. [1], the covering numbers in the above bound may be bounded in terms of the scale-sensitive dimension P_γ. They show that for all $\varepsilon\in(0,1)$,

$$\sup_{x_1,\ldots,x_n}N(\varepsilon,x_1^n,\mathscr{F})\leqslant 2\left(\frac{4n}{\varepsilon^2}\right)^{P_{\varepsilon/4}\log(2en/(P_{\varepsilon/4}))}. \tag{6}$$

This inequality was subsequently improved by Bartlett and Long [6].

Remark. Another method of exploiting the information $\eta^*\in\mathscr{F}$ is *maximum likelihood.* Here one selects a function from $\mathscr{F}$ by maximizing the likelihood

$$\prod_{i=1}^{n}\eta(X_i)^{Y_i}(1-\eta(X_i))^{1-Y_i}.$$

It is shown in [13, Ch. 15] that maximum likelihood works well whenever the *bracketing metric entropy* of $\mathscr{F}$ is finite.

3. Dimensions

So far, we have seen two different methods – one based on minimizing empirical misclassification and the other on empirical squared error minimization – such that one of them guarantees small probability of error if the VC dimension V is small, and the other when the scale-sensitive dimension P_ε is small. However, it is easy to see that no universal relationship exists between V and P_ε. In fact, if $\mathscr{F}$ is the class of all functions on $\mathscr{R}$ whose value is in $[0,\frac{1}{2}]$ if $x<0$ and in $(\frac{1}{2},1]$ if $x>0$ then $V=1$ but $P_\varepsilon=\infty$ for every ε. On the other hand, if $\mathscr{F}$ contains every function defined on the positive integers such that $|\eta(x)-\frac{1}{2}|\leqslant e^{-x}/2$, then $P_\varepsilon=\lfloor-\log(2\varepsilon)\rfloor$, but $V=\infty$. (The latter example is taken from [1].) Therefore, these two methods are not compatible. In fact, it is easy to show situations in which $V=\infty$, empirical error minimization fails

but empirical squared error minimization provides a small probability of error, and vice versa when $P_e^* = \infty$. The reason is that these general methods are not designed for the specific classification problem we are studying.

In this paper we look for methods which are universal in the sense that they unify the advantages of the above two principles, and in fact, may work well even if both previous methods fail. We discuss three different methods in detail. All of them are *skeleton estimates*, that is, first a finite subset of $\mathcal{F}$ is selected, and then the empirical error is minimized over this subset. The methods differ in the way the skeleton is formed. The first method – introduced in Section 4 – assumes the knowledge of the distribution μ of X. We show that the rate of convergence of this method is upper bounded by a quantity determined by the *minimum* of V and P_e^*. In Section 5 we discuss methods which do not assume the knowledge of μ. Here the skeleton is formed in a data-dependent way. The first simple method has performance guarantees better than those of empirical squared error minimization. The second data-based skeleton estimate is specifically suited for the classification problem. The size of the error here is dominated by the covering numbers according to a new metric, and the performance bounds we obtain here are better than those obtainable for empirical error minimization. In Theorem 4 we relate these covering numbers to a new scale-sensitive dimension d_γ defined below, which is always smaller than the minimum of V and P_γ.

Next, we define a dimension for a class $\mathcal{F}$ of functions $\mathcal{X} \to [0,1]$.

Definition 1. Let $0 < \gamma \leqslant \frac{1}{2}$. We say that $\mathcal{F}$ γ-shatters a finite set $A \subset \mathcal{X}$ if there exists some function $s : A \to [\frac{1}{2} - \gamma, \frac{1}{2} + \gamma]$ such that for every subset $E \subset A$ there is a function $\eta_E \in \mathcal{F}$ such that $\eta_E(x) \geqslant s(x) + \gamma$ if $x \in E$ and $\eta_E(x) < s(x) - \gamma$ if $x \in A - E$. The *γ-dimension* d_γ of $\mathcal{F}$ is defined as the largest positive integer n for which there exists a set A of cardinality n which is γ-shattered by $\mathcal{F}$. If for every n there is a set A which is γ-shattered by $\mathcal{F}$ then we say that $d_\gamma = \infty$.

Remark. The asymmetry in the definition of γ-shattering is necessary for our purposes. It corresponds to the tie-breaking convention we apply, that is, to the fact that when $\eta(x) = \frac{1}{2}$, the corresponding classification is always 1. If this asymmetry was not present, d_γ would be equal to the P_γ dimension of the set of "clipped" functions $\pi_{2\gamma}(\eta)$, $\eta \in \mathcal{F}$, where $\pi_{2\gamma}$ is defined for $y \in [0,1]$ as

$$\pi_{2\gamma}(y) = \begin{cases} \frac{1}{2} + 2\gamma & \text{if } y \geqslant \frac{1}{2} + 2\gamma, \\ \frac{1}{2} - 2\gamma & \text{if } y \leqslant \frac{1}{2} - 2\gamma, \\ y & \text{otherwise.} \end{cases}$$

This class of functions also appears in [4] in a somewhat different setup.

First note that d_γ is a monotone decreasing function of γ. To see this, let $\gamma_1 \geqslant \gamma_2$, and let A be a set which is γ_1-shattered by $\mathcal{F}$ according to some function

$s : A \to [\frac{1}{2} - \gamma_1, \frac{1}{2} + \gamma_1]$ Then it is easy to see that the same set is γ_2-shattered by $\mathscr{F}$ according to the function s' defined by

$$s'(x) = \begin{cases} s(x) & \text{if } s(x) \in [\frac{1}{2} - \gamma_2, \frac{1}{2} + \gamma_2], \\ \frac{1}{2} - \gamma_2 & \text{if } s(x) < \frac{1}{2} - \gamma_2, \\ \frac{1}{2} + \gamma_2 & \text{if } s(x) > \frac{1}{2} + \gamma_2. \end{cases}$$

Our main result concerning d_γ is Theorem 4, which is an upper bound for certain covering numbers appearing in the performance bound (Theorem 3) of a data-dependent skeleton estimate proposed in Section 5.

d_γ is closely related to the scale-sensitive dimension P_γ, whose usefulness have been demonstrated for learning "probabilistic concepts" and for more general regression function estimation problems, see [1, 2, 4, 6, 7, 17, 26]. The only difference is that in the definition of P_γ the range of the shattering function s is not restricted, it can take any value in $[0, 1]$. Therefore, clearly, for every γ,

$$d_\gamma \leqslant P_\gamma.$$

In fact, d_γ may be finite even if $P_\gamma = \infty$ for every γ. (Just consider the class of all functions $\eta : \mathscr{R} \to [0, 1]$ such that $\eta(x) < \frac{1}{2}$ if $x \leqslant 0$ and $\eta(x) \geqslant \frac{1}{2}$ if $x > 0$.) The restriction of the range of s is motivated by the fact that from the point of view of classification, only the behavior of the functions in $\mathscr{F}$ around $\frac{1}{2}$ matters. For some discussion on this we refer to Section 6.7 of [13].

Also clearly,

$$d_\gamma \leqslant V$$

for each γ. Again, d_γ may be finite even if $V = \infty$. As a simple example, consider the class of Lipschitz functions on $[0, 1]$. Then $d_\gamma \leqslant P_\gamma = O(1/\gamma)$, but obviously $V = \infty$.

Since $d_\gamma \leqslant \min(V, P_\gamma)$, we may interpret the new dimension as one that unifies the advantages of V and the scale-sensitive dimension P_γ. On the other hand, it is important to note that d_γ may be finite even if $\min(V, P_\gamma) = \infty$. The simplest example is the class of all functions $\mathscr{R} \to [0, \frac{1}{2} + \gamma/2]$.

Several other dimensions have been introduced to measure the size of classes of functions. A partial survey is found in [1].

4. Skeleton estimates

Skeleton estimates first form a finite subclass of the class of functions $\mathscr{F}$ and then use some kind of empirical risk minimization to choose an estimate from this class. Different types of skeleton estimates were proposed and studied by Benedek and Itai [8], Buescher and Kumar [11, 12], Dudley et al. [15], Kulkarni [18], Vapnik [27], and Devroye et al. [13, Ch. 28].

The "skeleton" is typically chosen as some kind of covering of the class $\mathscr{F}$, that is, for each function $\eta \in \mathscr{F}$, there should be a function in the skeleton which is sufficiently

close to η in some sense. Different versions of skeleton estimates differ in the way this closeness is measured, and also in the way the classifier is selected from the skeleton. For example, assuming that the distribution μ of X is known, Vapnik [27], and Benedek and Itai [8] cover the class of classifiers in the metric $d_H(\eta,\eta')=P\{g(X)\neq g'(X)\}$ and select by minimizing the empirical error over the covering. In [13, Ch. 28], $\mathscr{F}$ is covered with respect to the supremum norm $\sup_{x\in\mathscr{X}}|\eta(x)-\eta'(x)|$, and again, the empirical error is minimized over the covering.

In this paper we propose new skeleton estimates, specifically designed for the discussed classification problem, which perform better than previously discussed methods. The key ingredient of our method is a new way of selecting the skeleton. The "distance" according to which we measure closeness of members of $\mathscr{F}$ is motivated by (1).

For simplicity, assume first that the distribution of X is known. Then for $\varepsilon>0$, we may select a set of functions $\widehat{\mathscr{F}}_\varepsilon$ such that for every $\eta\in\mathscr{F}$ there is an $\eta'\in\widehat{\mathscr{F}}_\varepsilon$ such that

$$E\{2|\eta(X)-\eta'(X)|I_{\{g(X)\neq g'(X)\}}\}<\varepsilon.$$

(g and g' are the classifiers defined by η and η'.) Choose $\widehat{\mathscr{F}}_\varepsilon$ such that it has minimal cardinality. Next, select a member of $\widehat{\mathscr{F}}_\varepsilon$ which minimizes the empirical error

$$L_n(g)=\frac{1}{n}\sum_{i=1}^{n}I_{\{g(X_i)\neq Y_i\}}$$

over all $g=I_{\{\eta\geqslant 1/2\}}$, $\eta\in\widehat{\mathscr{F}}_\varepsilon$. Denote the minimizing function by $\widehat{\eta}_n$ and the corresponding classifier by $\widehat{g}_n$.

We have the following performance bound for this estimate:

Theorem 1. *If $\eta^*\in\mathscr{F}$, then for every $\varepsilon,\delta>0$,*

$$P\{L(\widehat{g}_n)-L^*>\delta+\varepsilon\}\leqslant(|\widehat{\mathscr{F}}_\varepsilon|+1)e^{-(3/32)n\delta^2/(L^*+\delta+\varepsilon)},$$

and

$$E\{L(\widehat{g}_n)\}-L^*\leqslant\varepsilon+\max\left[\sqrt{\frac{22(L^*+\varepsilon)\log(n(|\widehat{\mathscr{F}}_\varepsilon|+1))}{n}},\frac{22\log(n(|\widehat{\mathscr{F}}_\varepsilon|+1))}{n}\right].$$

All proofs are given in Section 6.

Observe that a straightforward consequence of the definition of the covering $\widehat{\mathscr{F}}_\varepsilon$ is that $|\widehat{\mathscr{F}}_\varepsilon|\leqslant\min(N_1(\varepsilon/2,\mu,\mathscr{F}),N_H(\varepsilon/2,\mu,\mathscr{F}))$. Therefore, the probability of error of the selected classifier is guaranteed to be small if *at least* one of these two covering numbers is controlled. With a bit of work, we may relate these covering numbers to the vc dimension and the Kearns–Shapire scale-sensitive dimension of $\mathscr{F}$. This implies the most important corollary of this theorem, which states that the obtained classifier has a small guaranteed probability of error whenever the *minimum* of the vc dimension V and the scale-sensitive dimension $P_{\varepsilon/256}$ is finite:

Corollary 1. *If* $\eta^* \in \mathscr{F}$, *then for every* $\varepsilon, \delta > 0$, *and* $\gamma > \varepsilon$ (*recall that* ε *is a parameter of the algorithm*),

$$P\{L(\hat{g}_n) - L^* > \gamma\} < \delta$$

for

$$n \geqslant K \frac{1}{\gamma^2} \left(\min(V, P_{\varepsilon/256}) \log^2 \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right).$$

The corollary shows that the rate of convergence of the proposed algorithm is determined by the *minimum* of the two dimensions that appeared in the two algorithms obtained by two different philosophies – namely, empirically minimizing misclassification and estimating regression functions. The skeleton estimate investigated here seems to unify the advantages of the two approaches. The constants and the power of the logarithmic factors have not been optimized. In this qualitative discussion we sacrifice sharp results for simplicity. The bound of Theorem 1 is better than those that can be obtained for the above-mentioned skeleton estimates of [27, 8] and [13, Ch. 28].

On the other hand, the algorithm may work even if $\min(V, P_\varepsilon) = \infty$ for every ε. Consider the example in which $\mathscr{F}$ contains all functions with values in $[0, \frac{1}{2} + \varepsilon/4]$. Then $\min(V, P_\varepsilon) = \infty$, and it is easy to see that both empirical error minimization and empirical squared error minimization may give huge probabilities of error. On the other hand, $|\hat{\mathscr{F}}_\varepsilon| = 1$, since the set $\mathscr{F}$ is covered by the single function $\eta \equiv \frac{1}{2} - (\varepsilon/4)$. We conjecture that $\min(V, P_{\varepsilon/256})$ in Corollary 1 can be replaced by $d_{K\varepsilon}$, but we have not been able to prove this.

The main disadvantage of the algorithm discussed here is that it assumes knowledge of the underlying distribution μ of X. In the next section we propose and analyze methods which do not use such information. It is interesting to note, however, that knowledge of μ usually does not help much. In fact, the minimax lower bound (3) remains valid even if μ is allowed to be known.

5. Data-based skeleton estimates

In this section we present classification rules that first form a finite skeleton of $\mathscr{F}$ based on a part of the training data, and then use the other half of the data to select the empirically best candidate from the skeleton. The idea of such empirical covering is due to Buescher and Kumar [12], and was further explored by Lugosi and Nobel [20]. The key difference here is that the metric according to which the covering is chosen is different from what is used in selecting a classifier from the covering. For example, in the simplest situation, one may first form an L_1-cover of $\mathscr{F}$, and then (instead of minimizing the empirical L_1 error) minimize the empirical misclassification over the skeleton. This hybrid approach seems to prove fruitful.

The precise definition of the classifier described above is as follows: first, the data sequence D_n is split into two parts:

$$D_m = ((X_1, Y_1), \ldots, (X_m, Y_m))$$

and

$$T_{n-m} = ((X_{m+1}, Y_{m+1}), \ldots, (X_n, Y_n)).$$

The first part, D_m, is used to create a skeleton of $\mathcal{F}$, that is, a finite set $\mathcal{F}_\varepsilon^\circ = \mathcal{F}_\varepsilon^\circ(D_m)$ such that each member of $\mathcal{F}$ is closely approximated by a function from $\mathcal{F}_\varepsilon^\circ$ in the empirical L_1 metric. More precisely, given $\varepsilon > 0$, let $\mathcal{F}_\varepsilon^\circ$ be a class of functions of minimal cardinality satisfying the property that for each $\eta \in \mathcal{F}$ there exists an $\bar{\eta} \in \mathcal{F}_\varepsilon^\circ$ such that

$$\frac{2}{m} \sum_{i=1}^{m} |\eta(X_i) - \bar{\eta}(X_i)| < \varepsilon.$$

Clearly, $|\mathcal{F}_\varepsilon^\circ| = N_1(\varepsilon/2, X_1^m, \mathcal{F})$. The second part of the data, T_{n-m}, is used to test all classifiers defined by functions in $\mathcal{F}_\varepsilon^\circ$, and to select one with minimal empirical error. In other words, we minimize the empirical error

$$L_{n-m}(g) = \frac{1}{n-m} \sum_{i=m+1}^{n} I_{\{g(X_i) \neq Y_i\}}$$

among all classifiers $g = I_{\{\eta \geqslant 1/2\}}$ in $\mathcal{F}_\varepsilon^\circ$. Denote the obtained classification rule by g_n° (and the corresponding regression function by η_n°). Note that g_n° ignores the values of $Y_1, \ldots, Y_m$, and therefore it may make efficient use of additional unlabeled samples, if available. The first half of the sample is only used to obtain information about μ. For simplicity, we only consider splitting the data into two equal halfs (i.e., $m = n/2$). This is a safe, but not necessarily optimal choice. The following performance bound for the obtained classifier is given for the case $m = n/2$, but the proof in Section 6 contains the general inequalities.

Theorem 2. *If $\eta^* \in \mathcal{F}$, then for any n, ε, and $\delta \geqslant 6\varepsilon$,*

$$P\{L(g_n^\circ) - L^* > 3\delta\} \leqslant 2EN_1(\varepsilon/2, X_1^{n/2}, \mathcal{F}) e^{-3n\delta^2/(32L^*)} + 10EN_1(\varepsilon/32, X_1^{n/2}, \mathcal{F}) e^{-n\delta/512}.$$

In particular,

$$E\{L(g_n^\circ)\} - L^* \leqslant 6\varepsilon + 3 \max\left[\sqrt{\frac{11L^* \log(2nEN_1(\varepsilon/2, X_1^{n/2}, \mathcal{F}))}{n}}, \frac{512 \log(10nEN_1(\varepsilon/32, X_1^{n/2}, \mathcal{F}))}{n}\right].$$

The bound obtained here is better than that of (5) for the method of squared error minimization in that it reflects the right dependence on L^* and that it involves the expected random covering numbers instead of their supremum. We believe that the

same type of bound is also true for squared error minimization and we doubt that the above skeleton estimate is inherently better than squared error minimization. We included this method because of its simplicity, conceptual reasons, and comparison purposes. Again, we may rephrase the upper probability inequality in terms of sample sizes and the scale-sensitive dimension P_γ. The proof of the next corollary is direct by applying (6):

Corollary 2. *If $\eta^* \in \mathscr{F}$, then for every $\varepsilon, \delta > 0$, and $\gamma > 18\varepsilon$ (recall that ε is a parameter of the algorithm),*

$$P\{L(g_n^\circ) - L^* > \gamma\} < \delta$$

for

$$n \geqslant K \max\left(\frac{L^* P_{\varepsilon/8}}{\gamma^2} \log^2 \frac{P_{\varepsilon/8}}{\varepsilon} + \frac{L^*}{\gamma^2} \log \frac{1}{\delta}, \frac{P_{\varepsilon/128}}{\gamma} \log^2 \frac{P_{\varepsilon/128}}{\varepsilon} + \frac{1}{\gamma} \log \frac{1}{\delta} \right).$$

Next, we apply the ideas of Section 4 to define a data-based skeleton estimate better suited for classification purposes. The main idea is to empirically cover $\mathscr{F}$ with respect to the empirical version of the "metric" $E\{2|\eta(X) - \eta'(X)| I_{\{g(X) \neq g'(X)\}}\}$. More precisely, given $\varepsilon > 0$, let $\overline{\mathscr{F}_\varepsilon}$ be a class of functions of minimal cardinality satisfying the property that for each $\eta \in \mathscr{F}$ there exists an $\bar{\eta} \in \overline{\mathscr{F}_\varepsilon}$ such that

$$\frac{2}{m} \sum_{i=1}^{m} |\eta(X_i) - \bar{\eta}(X_i)| I_{\{g(X_i) \neq \bar{g}(X_i)\}} < \varepsilon. \tag{7}$$

Then, just like in the case of the previous classifier, the empirical error

$$L_{n-m}(g) = \frac{1}{n-m} \sum_{i=m+1}^{n} I_{\{g(X_i) \neq Y_i\}}$$

is minimized among all classifiers $g = I_{\{\eta \geqslant 1/2\}}$ in $\overline{\mathscr{F}_\varepsilon}$. Denote the obtained classification rule and the corresponding regression function by $\bar{g}_n$ and $\bar{\eta}_n$. We have the following property:

Theorem 3. *If $\eta^* \in \mathscr{F}$, then for any n, ε, and $\delta \geqslant 6\varepsilon$,*

$$P\{L(\bar{g}_n) - L^* > 3\delta\} \leqslant 2E\{|\overline{\mathscr{F}_\varepsilon}|\} e^{-3n\delta^2/(32L^*)} + 2E\{|\overline{\mathscr{F}_\varepsilon}|\} e^{-n\delta/32} + 8E N_1(\varepsilon/16, X_1^{n/2}, \mathscr{G}) e^{-n\delta/512},$$

where $\mathscr{G}$ is the class of functions $|2\eta^*(x) - 1| I_{\{g(x) \neq g^*(x)\}}$, $\eta \in \mathscr{F}$. In particular,

$$E\{L(\bar{g}_n)\} - L^* \leqslant 6\varepsilon + 3 \max\left[\sqrt{\frac{11L^* \log(2nE\{|\overline{\mathscr{F}_\varepsilon}|\})}{n}}, \frac{512 \log(10n \max[E\{|\overline{\mathscr{F}_\varepsilon}|\}, E N_1(\varepsilon/16, X_1^{n/2}, \mathscr{G})])}{n} \right].$$

Thus, the rate of convergence of the error of the selected classifier is determined by the logarithm of the expected value of the covering number $|\overline{\mathscr{F}_\varepsilon}|$. The lower-order term involves the expected L_1 covering numbers of the class $\mathscr{G}$. Unfortunately, this covering number cannot be bounded by $|\overline{\mathscr{F}_\varepsilon}|$, and in fact, in some situations it can be much larger. However, note that for any $\eta, \eta' \in \mathscr{F}$,

$$\begin{aligned}&||2\eta^*(x)-1|I_{\{g(x)\neq g^*(x)\}}-|2\eta^*(x)-1|I_{\{g'(x)\neq g^*(x)\}}|\\&\quad=I_{\{g(x)\neq g'(x)\}}|2\eta^*(x)-1|\leqslant I_{\{g(x)\neq g'(x)\}}.\end{aligned}$$

Therefore, $N_1(\varepsilon/16, X_1^{n/2}, \mathscr{G})) \leqslant N_H(\varepsilon/16, X_1^{n/2}, \mathscr{F}))$, and in particular, for $\varepsilon \geqslant 32/n$,

$$N_1(\varepsilon/16, X_1^{n/2}, \mathscr{G})) \leqslant S_{\mathscr{F}}(X_1^{n/2}) \leqslant \left(\frac{ne}{2V}\right)^V. \tag{8}$$

Thus, the finiteness of the VC dimension and $L^* > 0$ guarantees that for sufficiently large sample sizes the maximum is taken by the first term, involving $E\{|\overline{\mathscr{F}_\varepsilon}|\}$. If $V = \infty$, we cannot guarantee any nontrivial bound for $E\{L(\bar{g}_n)\}$ (see also the remark below).

However, since $E\{N_1(\varepsilon/16, X_1^{n/2}, \mathscr{G})\}$ only appears in the lower-order term, in most cases it has a minor importance. The main message is that the expected size of the error is of the order of $\sqrt{L^* \log(E\{|\overline{\mathscr{F}_\varepsilon}|\})/n}$, unless the Bayes error L^* is very small.

Below in Theorem 4 we relate the new scale-sensitive dimension d_γ, introduced in Section 3, to the covering numbers appearing in Theorem 3.

It is worth comparing $E\{|\overline{\mathscr{F}_\varepsilon}|\}$ to other quantities relevant for analyzing the performance of different classification rules. For example, it is easy to see that for all $\varepsilon \geqslant 4/n$,

$$E\{|\overline{\mathscr{F}_\varepsilon}|\} \leqslant E\{S_{\mathscr{F}}(X_1^{n/2})\}.$$

(Note that "good" choices for ε are always larger, and in some cases much larger than $4/n$.) Thus, for the data-dependent skeleton classifier, Theorem 3 guarantees better performance than the inequality (2) for empirical risk minimization. The reason why it is possible to improve on these bounds is that one can make use of the additional information provided by the knowledge of the form of the possible regression functions.

We must remark here, however, that in order to obtain Theorem 3 (as well as Theorems 1, 2), we needed to assume that $\eta^* \in \mathscr{F}$, that is, the "true" regression function is the member of the class $\mathscr{F}$. The bounds obtained for empirical risk minimization do not require this assumption. Empirical risk minimization is, therefore, much more robust than the skeleton estimate introduced here.

It is also easy to see that

$$E\{|\overline{\mathscr{F}_\varepsilon}|\} \leqslant E\{N_1(\varepsilon/2, X_1^{n/2}, \mathscr{F})\}.$$

Therefore, the rate of convergence achieved by the classifier $\bar{g}_n$ is usually better than that of empirical squared error minimization (or than that of the first data-dependent skeleton classifier introduced in this section). Consider, for example, the following trivial case: let $\mathscr{F}$ contain *all* functions $\eta : \mathscr{R} \to [0,1]$ such that $\eta(x) < \frac{1}{2}$ if $x < 0$ and

$\eta(x) \geq \frac{1}{2}$ if $x \geq 0$. Then there is only one classifier induced by these functions, and accordingly, $|\overline{\mathscr{F}_\varepsilon}| = 1$. On the other hand, this class is clearly too large for obtaining meaningful bounds for regression function estimation.

Remark. Note, however, that $\bar{g}_n$ may fail in some situations when empirical squared error minimization and g_n° do not. We have the following counterexample: Let $\mathscr{F}$ contain all functions on $[0,1]$ which take their values between $\frac{1}{2} - \varepsilon/4$ and $\frac{1}{2} + \varepsilon/4$, and the constant zero function. Assume that $\eta^*(x) = 0$ for all $x \in [0,1]$. Then $|\overline{\mathscr{F}_\varepsilon}| = 1$, since $\mathscr{F}$ can be covered by the single function which equals $\frac{1}{2} - \varepsilon/4$ if x is one of the first $n/2$ data points, and $\frac{1}{2} + \varepsilon/4$ otherwise. If the distribution of X is absolutely continuous, $L(\bar{g}_n) = 1$ with probability one, while $L^* = 0$. On the other hand, $EN_1(\varepsilon/32, X_1^{n/2}, \mathscr{F})) = 2$, so, for example, g° is guaranteed to give a small error probability.

Next, we relate $|\overline{\mathscr{F}_\varepsilon}|$ to the new scale-sensitive dimension d_γ defined in Section 3.

Theorem 4. *For any value of* $X_1, \ldots, X_m$, *if* $12m \geq \lceil 1/\varepsilon \rceil$, $\varepsilon \leq 1$, *then*

$$|\overline{\mathscr{F}_\varepsilon}| \leq 2(6m)^{d_{\varepsilon/10} \log_2(2em/d_{\varepsilon/10})+1}.$$

This result is the analogue of (6) for the new dimension introduced here. The proof is based on the proof of Lemma 3.4 of Alon et al. [1], and it is given in Section 6.

We may combine the above result in a straightforward way with Theorem 3 and (8) to obtain the following sample-size bound for the data-dependent skeleton estimate $\bar{g}_n$ defined in Section 5.

Corollary 3. *Assume that* $\eta^* \in \mathscr{F}$. *For every* $\varepsilon, \delta > 0$, *and* $\gamma > 18\varepsilon$ (*recall that* ε *is a parameter of the algorithm*),

$$P\{L(\bar{g}_n) - L^* > \gamma\} < \delta$$

for

$$K \max\left(\frac{L^* d_{\varepsilon/20}}{\gamma^2} \log^2 \frac{d_{\varepsilon/20}}{\varepsilon} + \frac{L^*}{\gamma^2} \log \frac{1}{\delta}, \frac{V}{\gamma} \log^2 \frac{V}{\varepsilon} + \frac{1}{\gamma} \log \frac{1}{\delta} \right).$$

Therefore, if $V < \infty$ and $L^* > 0$, for sufficiently small γ's the sample size is determined by $d_{\varepsilon/20}$ instead of the VC dimension as for empirical error minimization.

Finally, we indicate some points where the bound of Theorem 4 is loose. First of all, the theorem provides an upper bound for the maximal possible value of $|\overline{\mathscr{F}_\varepsilon}|$, whereas the interesting quantity in Theorem 3 is its expected value $E\{|\overline{\mathscr{F}_\varepsilon}|\}$. In certain cases, the difference may be significant: [13, Theorem 13.13] provides such an example.

If $\mathscr{F}$ is a class of indicator functions, then $|\overline{\mathscr{F}_\varepsilon}|$ is just the random shatter coefficient $S_{\mathscr{F}}(X_1^m)$ for $\varepsilon = 2/m$. In this case Sauer's lemma [25] implies that $\log |\overline{\mathscr{F}_\varepsilon}| \leq V \log m$, whereas Theorem 4 only gives $\log |\overline{\mathscr{F}_\varepsilon}| = O(V \log^2 m)$.

If $\mathscr{F}$ is the class of all Lipschitz functions (with Lipschitz constant 1) on $[0,1]$, then it is easy to see that $\log|\overline{\mathscr{F}_\varepsilon}| = O(1/\varepsilon)$. However, it is also easy to see that $d_\varepsilon = O(1/\varepsilon)$, and therefore Theorem 4 only implies $\log|\overline{\mathscr{F}_\varepsilon}| = O(\log^2 m/\varepsilon)$. However, the practical importance of the log factors is minor, so Theorem 4 may be a useful tool to bound $|\overline{\mathscr{F}_\varepsilon}|$.

6. Proofs

A key ingredient of several proofs is the following simple lemma:

Lemma 1. *Let $\mathscr{F}'$ be a finite set of functions $\mathscr{X} \to [0,1]$, possibly depending on D_m. Let η_n minimize the empirical error*

$$L_{n-m}(\eta) = \frac{1}{n-m}\sum_{i=m+1}^{n} I_{\{g(X_i)\neq Y_i\}}$$

over $\eta \in \mathscr{F}'$ and let $\widehat{\eta}$ be minimize the probability of error $L(\widehat{\eta})$ in $\mathscr{F}'$. Let g_n and $\widehat{g}$ denote the corresponding classifiers. Then

$$P\{L(g_n) - L(\widehat{\eta}) > \delta | D_m\} \leqslant (|\mathscr{F}'| + 1)e^{-(3/8)(n-m)\delta^2/(L(\widehat{\eta})+2\delta)}. \tag{9}$$

Proof. If $L(g_n) - L(\widehat{\eta}) > 2\delta$, then there exists an $\eta \in \mathscr{F}'$ such that $L(\eta) > L(\widehat{\eta}) + 2\delta$ and $L_{n-m}(\eta) \leqslant L_{n-m}(\widehat{\eta})$. Thus,

$$\begin{aligned}
&P\{L(g_n) - L(\widehat{\eta}) > 2\delta | D_m\} \\
&\quad \leqslant P\left\{ \min_{\eta: L(\eta) > L(\widehat{\eta})+2\delta} L_{n-m}(\eta) < L_{n-m}(\widehat{\eta}) \,\middle|\, D_m \right\} \\
&\quad \leqslant P\left\{ \min_{\eta: L(\eta) > L(\widehat{\eta})+2\delta} L_{n-m}(\eta) < L(\widehat{\eta}) + \delta \,\middle|\, D_m \right\} \\
&\qquad + P\{L_{n-m}(\widehat{\eta}) > L(\widehat{\eta}) + \delta | D_m\},
\end{aligned}$$

so we need to show that

$$\begin{aligned}
&P\left\{ \min_{\eta: L(\eta) > L(\widehat{\eta})+2\delta} L_{n-m}(\eta) < L(\widehat{\eta}) + \delta \,\middle|\, D_m \right\} \\
&\quad \leqslant P\left\{ \max_{\eta \in \mathscr{F}'} \frac{L(\eta) - L_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta}) + 2\delta}} \,\middle|\, D_m \right\}.
\end{aligned} \tag{10}$$

But if

$$\max_{\eta \in \mathscr{F}'} \frac{L(\eta) - L_{n-m}(\eta)}{\sqrt{L(\eta)}} \leqslant \frac{\delta}{\sqrt{L(\widehat{\eta}) + 2\delta}},$$

then for each $\eta \in \mathscr{F}'$

$$L_{n-m}(\eta) \geqslant L(\eta) - \delta\sqrt{\frac{L(\eta)}{L(\widehat{\eta}) + 2\delta}}.$$

If, in addition, η is such that $L(\eta) > L(\widehat{\eta}) + 2\delta$, then by the monotonicity of the function $x - c\sqrt{x}$ (for $c > 0$ and $x > c^2/4$),

$$L_{n-m}(\eta) \geqslant L(\widehat{\eta}) + 2\delta - \delta\sqrt{\frac{L(\widehat{\eta}) + 2\delta}{L(\widehat{\eta}) + 2\delta}} = L(\widehat{\eta}) + \delta,$$

and (10) follows. Thus, we have

$$P\{L(g_n) - L(\widehat{\eta}) > 2\delta | D_m\} \leqslant P\{L_{n-m}(\widehat{\eta}) - L(\widehat{\eta}) > \delta | D_m\} + P\left\{\max_{\eta \in \mathscr{F}_c} \frac{L(\eta) - L_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta}) + 2\delta}} \middle| D_m\right\}.$$

Next, we bound both terms on the right-hand side of the above inequality. First, since given D_m, the conditional distribution of $(n-m)L_{n-m}(\widehat{\eta})$ is binomial with parameters $n-m$ and $L(\widehat{\eta})$, we have by Bernstein's inequality [9] that

$$P\{L_{n-m}(\widehat{\eta}) - L(\widehat{\eta}) > \delta | D_m\} \leqslant \mathrm{e}^{-(n-m)\delta^2/(2L(\widehat{\eta}) + (2/3)\delta)}.$$

(Recall that $\widehat{\eta}$ minimizes the probability of error in $\mathscr{F}'$.) On the other hand, clearly

$$P\left\{\max_{\eta \in \mathscr{F}'} \frac{L(\eta) - L_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta}) + 2\delta}} \middle| D_m\right\} \leqslant |\mathscr{F}'| \max_{\eta \in \mathscr{F}'} P\left\{\frac{L(\eta) - L_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta}) + 2\delta}} \middle| D_m\right\}.$$

For any fixed η, the probability on the right-hand side is zero if $\gamma \stackrel{\text{def}}{=} \delta/\sqrt{L(\widehat{\eta}) + 2\delta} > \sqrt{L(\eta)}$. Otherwise, if $\gamma \leqslant \sqrt{L(\eta)}$, then again by Bernstein's inequality,

$$P\left\{\frac{L(\eta) - L_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta}) + 2\delta}} \middle| D_m\right\}$$

$$= P\{L(\eta) - L_{n-m}(\eta) > \gamma\sqrt{L(\eta)} \,|\, D_m\} \leqslant \mathrm{e}^{-\frac{(n-m)\gamma^2 L(\eta)}{2L(\eta) + 2\sqrt{L(\eta)}\gamma/3}}$$

$$\leqslant \mathrm{e}^{-(3/8)(n-m)\gamma^2}$$

$$= \mathrm{e}^{-(3/8)(n-m)\delta^2/(L(\widehat{\eta}) + 2\delta)},$$

which completes the proof of the lemma. □

We have the following easy corollary:

Corollary 4.

$$E\{L(g_n) - L(\hat{\eta})|D_m\}$$
$$\leqslant \max\left[\sqrt{\frac{22L(\hat{\eta})\log(n(|\mathscr{F}'|+1))+1}{n-m}}, \frac{22\log(n(|\mathscr{F}'|+1))+1}{n-m}\right].$$

Proof. Observe that Lemma 1 implies that for all $\delta > 0$,

$$P\{L(g_n) - L(\hat{\eta}) > \delta | D_m\} \leqslant (|\mathscr{F}'|+1)\max[e^{-(3/64)(n-m)\delta^2/L(\hat{\eta})}, e^{-(3/64)(n-m)\delta}].$$

But for every u,

$$E\{L(g_n) - L(\hat{\eta})|D_m\} \leqslant u + P\{L(g_n) - L(\hat{\eta}) > u|D_m\}.$$

Choosing

$$u = \max\left[\sqrt{\frac{64L(\hat{\eta})\log(n(|\mathscr{F}'|+1))}{3(n-m)}}, \frac{64\log(n(|\mathscr{F}'|+1))+1}{3(n-m)}\right]$$

yields the corollary. □

Proof of Theorem 1. Note that by the inequality (1),

$$\min_{\eta\in\widehat{\mathscr{F}}_\varepsilon} L(\eta) - L^* \leqslant \varepsilon.$$

The Theorem now is a direct consequence of Lemma 1 and Corollary 4 (by taking $m=0$). □

Proof of Corollary 1. As we noted it before, $|\widehat{\mathscr{F}}_\varepsilon| \leqslant \min(N_1(\varepsilon/2,\mu,\mathscr{F}), N_H(\varepsilon/2,\mu,\mathscr{F}))$. The basic idea is that whenever deviations of empirical averages from their theoretical counterparts are uniformly small, empirical covering numbers are, in some sense, close to covering numbers measured by the corresponding expected distance. (This idea is quite standard, see [30, 22, 5].) Clearly, if

$$\sup_{\eta,\eta'\in\mathscr{F}} d_H(\eta,\eta') - d_{H,n}(\eta,\eta') < \frac{\varepsilon}{4},$$

then $d_H(\eta,\eta') < \varepsilon/2$ for every pair $\eta,\eta'\in\mathscr{F}$ such that $d_{H,n}(\eta,\eta') < \varepsilon/4$. This implies that

$$P\{N_H(\varepsilon/4, X_1^n, \mathscr{F}) < N_H(\varepsilon/2,\mu,\mathscr{F})\}$$
$$\leqslant P\left\{\sup_{\eta,\eta'\in\mathscr{F}} d_H(\eta,\eta') - d_{H,n}(\eta,\eta') \geqslant \frac{\varepsilon}{4}\right\}$$
$$\leqslant 4ES_{\mathscr{F}\cdot\mathscr{F}}(X_1^{2n})e^{-n\varepsilon^2/128}$$

(by the Vapnik–Chervonenkis inequality [28]),

where $\mathscr{F}-\mathscr{F}$ denotes class of functions $I_{\{g(x)\neq g'(x)\}}$, $\eta,\eta'\in\mathscr{F}$. But it is easy to see that, for example, $S_{\mathscr{F}-\mathscr{F}}(x_1^{2n})\leqslant S_{\mathscr{F}}(x_1^{2n})^4$, so we have that

$$P\{N_H(\varepsilon/4,X_1^n,\mathscr{F})<N_H(\varepsilon/2,\mu,\mathscr{F})\}\leqslant 4\left(\frac{2ne}{V}\right)^{4V}\mathrm{e}^{-n\varepsilon^2/128}.$$

The right-hand side of the inequality is strictly smaller than one if n is larger than a constant times $(V/\varepsilon^2)\log(1/\varepsilon)$. For such n's, there exist n points $x_1,\ldots,x_n$ such that $N_H(\varepsilon/4,x_1^n,\mathscr{F})\geqslant N_H(\varepsilon/2,\mu,\mathscr{F})$. But by Sauer's lemma, if $\varepsilon\geqslant 4/n$,

$$\sup_{x_1,\ldots,x_n} N_H(\varepsilon/4,x_1^n,\mathscr{F})\leqslant \sup_{x_1,\ldots,x_n} S_{\mathscr{F}}(x_1^n)\leqslant\left(\frac{ne}{V}\right)^V,$$

which means that for such n, $|\widehat{\mathscr{F}}_\varepsilon|\leqslant N_H(\varepsilon/2,\mu,\mathscr{F})\leqslant(ne/V)^V$.

To obtain an analogous upper bound for $|\widehat{\mathscr{F}}_\varepsilon|$ in terms of the scale-sensitive dimension P_γ, we proceed similarly:

$$\begin{aligned}P\{N_1(\varepsilon/4,X_1^n,\mathscr{F})<N_1(\varepsilon/2,\mu,\mathscr{F})\}&\leqslant P\left\{\sup_{\eta,\eta'\in\mathscr{F}} d_1(\eta,\eta')-d_{1,n}(\eta,\eta')\geqslant\frac{\varepsilon}{4}\right\}\\&\leqslant 8EN_1(\varepsilon/32,X_1^n,\mathscr{F}-\mathscr{F})\mathrm{e}^{-n\varepsilon^2/2048}\\&\qquad\text{(by Pollard [23, Ch. II])},\end{aligned}$$

where $\mathscr{F}-\mathscr{F}$ denotes the class of functions $\eta(x)-\eta'(x)$, $\eta,\eta'\in\mathscr{F}$. But an elementary argument shows that

$$N_1(\varepsilon/32,X_1^n,\mathscr{F}-\mathscr{F})\leqslant N_1(\varepsilon/64,X_1^n,\mathscr{F})^2.$$

We may bound these covering numbers by the key result (6), according to which

$$P\{N_1(\varepsilon/4,X_1^n,\mathscr{F})<N_1(\varepsilon/2,\mu,\mathscr{F})\}\leqslant 4\left(\frac{2^{14}n}{\varepsilon^2}\right)^{2P_{\varepsilon/256}\log(2en/(P_{\varepsilon/256}))}\mathrm{e}^{-n\varepsilon^2/2048}.$$

The right-hand side is strictly smaller than one if n is larger than a constant times

$$\frac{P_{\varepsilon/256}}{\varepsilon^2}\log^2\frac{1}{\varepsilon}.$$

For such n's therefore we have that

$$\begin{aligned}|\widehat{\mathscr{F}}_\varepsilon|&\leqslant N_1(\varepsilon/2,\mu,\mathscr{F})\\&\leqslant\sup_{x_1,\ldots,x_n} N_1(\varepsilon/4,x_1^n,\mathscr{F})\\&\leqslant 2\left(\frac{2^{14}n}{\varepsilon^2}\right)^{2P_{\varepsilon/256}\log(2en/(P_{\varepsilon/256}))},\end{aligned}$$

where again we used (6).

But Theorem 1 implies that for $\gamma > \varepsilon$,

$$P\{L(\widehat{g}_n) - L^* > \gamma\} < \delta \quad \text{if} \quad n \geqslant K \frac{1}{\gamma^2}\left(\log|\widehat{\mathscr{F}}_\varepsilon| + \log\frac{1}{\delta}\right),$$

so by putting the pieces together, the proof of the Corollary is finished. □

In the proof of Theorems 2 and 3 we apply an inequality of Pollard [24], sharpened by Haussler [16]. In particular, the following corollary is used, which was obtained by Buescher and Kumar [12] in a slightly different form. The form given here is found in Lugosi and Nobel [20].

Lemma 2. *Let $\mathscr{H}$ be class of functions on $\mathscr{X}$ such that $h(x) \in [0, A]$ for every $h \in \mathscr{H}$ and every $x \in \mathscr{X}$. Let $X_1, \ldots, X_m \in \mathscr{X}$ be i.i.d. random vectors. Then for each $\delta > 0$ and, $\varepsilon > 0$,*

$$P\left\{\sup_{h \in \mathscr{H}: (1/m)\sum_{i=1}^m h(X_i) < \varepsilon} E\{h(X)\} > \delta + 3\varepsilon\right\} \leqslant 4E\{|\mathscr{H}_{\varepsilon/16}|\}\mathrm{e}^{-m(\delta+\varepsilon)/(64A)},$$

where $\mathscr{H}_\varepsilon$ is any set of functions satisfying the property that for each $h \in \mathscr{H}$ there is a $h' \in \mathscr{H}_\varepsilon$ with

$$\frac{1}{m}\sum_{i=1}^m |h(X_i) - h'(X_i)| < \varepsilon.$$

Proof of Theorem 2. Write

$$L(g_n^\circ) - L^* = \left(L(g_n^\circ) - \min_{\eta \in \mathscr{F}_\varepsilon^\circ} L(\eta)\right) + \left(\min_{\eta \in \mathscr{F}_\varepsilon^\circ} L(\eta) - L^*\right),$$

so that

$$P\{L(g_n^\circ) - L^* > 3\delta\} \leqslant P\left\{L(g_n^\circ) - \min_{\eta \in \mathscr{F}_\varepsilon^\circ} L(\eta) > 2\delta\right\} + P\left\{\min_{\eta \in \mathscr{F}_\varepsilon^\circ} L(\eta) - L^* > \delta\right\}. \tag{11}$$

First we bound the second probability. Introduce the notation $h_\eta(x) = 2|\eta(x) - \eta^*(x)|$,

$$J_m(\eta) = \frac{1}{m}\sum_{i=1}^m h_\eta(X_i) \quad \text{and} \quad J(\eta) = E\{J_m(\eta)\} = E\{h_\eta(X)\}$$

for all $\eta \in \mathscr{F}$. Recall that by (1), $L(\eta) - L^* \leqslant J(\eta)$, and that $\min_{\eta \in \mathscr{F}_\varepsilon^\circ} J_m(\eta) < \varepsilon$ by the definition of $\mathscr{F}_\varepsilon^\circ$ and by the assumption $\eta^* \in \mathscr{F}$. Therefore,

$$\min_{\eta \in \mathscr{F}_\varepsilon^\circ} L(\eta) - L^* \leqslant \min_{\eta \in \mathscr{F}_\varepsilon^\circ} J(\eta) \leqslant \sup_{\eta \in \mathscr{F}: J_m(\eta) < \varepsilon} J(\eta).$$

Therefore, for $\delta \geqslant 6\varepsilon$,

$$P\left\{\min_{\eta\in\mathscr{F}_\varepsilon^\circ} L(\eta) - L^* > \delta\right\} \leqslant P\left\{\min_{\eta\in\mathscr{F}_\varepsilon^\circ} L(\eta) - L^* > \frac{\delta}{2} + 3\varepsilon\right\}$$

$$\leqslant P\left\{\sup_{\eta\in\mathscr{F}: J_m(\eta)<\varepsilon} J(\eta) > \frac{\delta}{2} + 3\varepsilon\right\}$$

$$\leqslant 4E\{N_1(\varepsilon/16, X_1^m, \mathscr{H})\}\, \mathrm{e}^{-m\delta/256} \tag{12}$$

by Lemma 2, where $\mathscr{H}$ denotes the class of functions h_η, $\eta \in \mathscr{F}$. By noting that for every η, η', and x, $|h_\eta(x) - h_{\eta'}(x)| \leqslant 2|\eta(x) - \eta'(x)|$, we see that $N_1(\varepsilon/16, X_1^m, \mathscr{H}) \leqslant N_1(\varepsilon/32, X_1^m, \mathscr{F})$, so for $\delta \geqslant 6\varepsilon$ we have

$$P\left\{\min_{\eta\in\mathscr{F}_\varepsilon^\circ} L(\eta) - L^* > \delta\right\} \leqslant 4E\{N_1(\varepsilon/32, X_1^m, \mathscr{F})\}\mathrm{e}^{-m\delta/256}.$$

The first probability on the right-hand side of (11) may directly be bounded by applying Lemma 1. Summarizing, we get

$$P\{L(g_n^\circ) - L^* > 3\delta\}$$

$$\leqslant P\left\{L(g_n^\circ) - \min_{\eta\in\mathscr{F}_\varepsilon^\circ} L(\eta) > 2\delta\right\} + P\left\{\min_{\eta\in\mathscr{F}_\varepsilon^\circ} L(\eta) - L^* > \delta\right\}$$

$$\leqslant P\left\{L(g_n^\circ) - \min_{\eta\in\mathscr{F}_\varepsilon^\circ} L(\eta) > 2\delta \,\middle|\, \min_{\eta\in\mathscr{F}_\varepsilon^\circ} L(\eta) - L^* \leqslant \delta\right\}$$

$$+ 2P\left\{\min_{\eta\in\mathscr{F}_\varepsilon^\circ} L(\eta) - L^* > \delta\right\}$$

$$\leqslant 2E\{N_1(\varepsilon/2, X_1^m, \mathscr{F})\}\mathrm{e}^{-(3/8)(n-m)\delta^2/(L^*+3\delta)} + 8E\{N_1(\varepsilon/32, X_1^m, \mathscr{F})\}\mathrm{e}^{-m\delta/256}$$

(by Lemma 1)

$$\leqslant 2E\{N_1(\varepsilon/2, X_1^m, \mathscr{F})\}\max[\mathrm{e}^{-(3/16)(n-m)\delta^2/L^*}, \mathrm{e}^{-(n-m)\delta/16}]$$

$$+ 8E\{N_1(\varepsilon/32, X_1^m, \mathscr{F})\}\mathrm{e}^{-m\delta/256}$$

$$\leqslant 2E\{N_1(\varepsilon/2, X_1^m, \mathscr{F})\}\mathrm{e}^{-(3/16)(n-m)\delta^2/L^*} + 10E\{N_1(\varepsilon/32, X_1^m, \mathscr{F})\}\mathrm{e}^{-m\delta/256},$$

so the proof of the probability inequality is finished. To obtain the bound for the expected value of the probability of error, simply observe that for any $u \geqslant 6\varepsilon$,

$$E\{L(g_n^\circ)\} - L^*$$

$$\leqslant 3u + P\{L(g_n^\circ) - L^* > 3u\}$$

$$\leqslant 3u + 2E\{N_1(\varepsilon/2, X_1^m, \mathscr{F})\}\mathrm{e}^{-(3/16)(n-m)u^2/L^*} + 10E\{N_1(\varepsilon/32, X_1^m, \mathscr{F})\}\mathrm{e}^{-mu/256},$$

and recalling that $m = n/2$, choose

$$u = 6\varepsilon + \max\left[\sqrt{\frac{11L^* \log(2nEN_1(\varepsilon/2, X_1^{n/2}, \mathscr{F}))}{n}}, \frac{512 \log(10nEN_1(\varepsilon/32, X_1^{n/2}, \mathscr{F}))}{n}\right]. \quad \square$$

Proof of Theorem 3. The proof is completely analogous to that of Theorem 2, so we omit it. □

In the rest of the paper we give the proof of Theorem 4.

The line of the proof of Theorem 4 is analogous to that of Lemma 3.4 in [1]. Just like there, we also begin with "discretizing". First, we introduce a discrete analogue of the γ-dimension, related to the "strong dimension" of [1].

Let b be a positive even integer, and let $\mathscr{G}$ be a class of functions $\mathscr{X} \to \{1, 2, \ldots, b\}$. We say that $\mathscr{G}$ *b-shatters* a finite set $A \subset \mathscr{X}$ according to a function $s : A \to \{b/2, b/2 + 1\}$ if for every subset E of A, there exists a function $f_E \in \mathscr{G}$ such that

$$f_E(x) \begin{cases} \leqslant s(x) - 1 & \text{if } x \in E, \\ \geqslant s(x) + 1 & \text{if } x \in A - E. \end{cases}$$

We say that $\mathscr{G}$ *b-shatters A* if $\mathscr{G}$ *b-shatters A* according to some s. The *b-dimension* Δ_b of $\mathscr{G}$ is the largest integer n such that there exists a set A, shattered by $\mathscr{G}$, with $|A| = n$. If there is no such larger integer, then we say that $\Delta_b = \infty$.

Let $\gamma > 0$. The γ-discretization of a function $\eta : \mathscr{X} \to [0, 1]$ is defined by $\eta^\gamma(x) = \text{round}(\eta(x)/\gamma) + 1$, where

$$\text{round}(z) = \begin{cases} \lfloor z \rfloor & \text{if } z - \lfloor z \rfloor < \frac{1}{2}, \\ \lfloor z \rfloor + 1 & \text{if } z - \lfloor z \rfloor \geqslant \frac{1}{2}. \end{cases}$$

Note that with $\gamma = 1/(b-1)$, and $\eta \in [0, 1]$, the γ-discretization $\eta^\gamma(x)$ of η maps $\mathscr{X}$ to $\{1, \ldots, b\}$.

Lemma 3. *Assume that $\gamma = 1/(b-1)$ for some positive even integer b, and let $\mathscr{F}^\gamma = \{\eta^\gamma : \eta \in \mathscr{F}\}$ denote the class of γ-discretizations of functions in $\mathscr{F}$. If Δ_b denotes the b-dimension of $\mathscr{F}^\gamma$, then*

$$\Delta_b \leqslant d_{\gamma/2},$$

where $d_{\gamma/2}$ is defined in Definition 1 for $\mathscr{F}$.

Proof. We show that if $\mathscr{F}^\gamma$ b-shatters a set A, then $\mathscr{F}$ $\gamma/2$-shatters A. Let $s : A \to \{b/2, b/2 + 1\}$ be the function which is used by $\mathscr{F}^\gamma$ to b-shatter A. Then for every

$E \subset A$ there is a function $\eta_E \in \mathscr{F}$ such that

$$\eta_E^{\gamma}(x) \begin{cases} \leqslant s(x) - 1 & \text{if } x \in E, \\ \geqslant s(x) + 1 & \text{if } x \in A - E. \end{cases}$$

Then clearly, if $s'(x) = \gamma s(x) - \gamma$, then

$$\eta_E(x) \begin{cases} < s'(x) - \gamma/2 & \text{if } x \in E, \\ \geqslant s'(x) + \gamma/2 & \text{if } x \in A - E, \end{cases}$$

so $\mathscr{F}$ clearly $\gamma/2$-shatters A. □

Next, we relate $|\overline{\mathscr{F}}_\varepsilon|$ to certain packing numbers of the class of discretizations of functions in $\mathscr{F}$. Let $\{x_1, \ldots, x_m\} \subset \mathscr{X}$. We say that a subset $\mathscr{F}'$ of $\mathscr{F}$ is *ε-separated* if for any $\eta_1, \eta_2 \in \mathscr{F}'$,

$$\frac{1}{m} \sum_{i=1}^{m} 2I_{\{g_1(x_i) \neq g_2(x_i)\}} |\eta_1(x_i) - \eta_2(x_i)| \geqslant \varepsilon.$$

The maximal size $M(\varepsilon, \mathscr{F})$ of such an ε-separated set is called the *ε-packing number* of $\mathscr{F}$. Now consider a class $\mathscr{G}$ of functions $\mathscr{X} \to \{1, \ldots, b\}$, where b is an even positive integer. We say that $\mathscr{G}' \subset \mathscr{G}$ is 2-*separated* if for any $f_1, f_2 \in \mathscr{G}'$,

$$\max_{i=1,\ldots,m} |f_1(x_i) - f_2(x_i)| I_{\{u(f_1(x_i)) \neq u(f_2(x_i))\}} \geqslant 2,$$

where the function u is defined by

$$u(a) = \begin{cases} 1 & \text{if } a > b/2, \\ 0 & \text{if } a \leqslant b/2. \end{cases}$$

The maximal size of a 2-separated subset of $\mathscr{G}$ is denoted by $M_b(2, \mathscr{G})$. The proof of the next lemma is trivial.

Lemma 4. *Let b be a positive even integer, and assume that $\gamma = 1/(b-1) \leqslant \varepsilon/4$. Then*

$$|\overline{\mathscr{F}}_\varepsilon| \leqslant M(\varepsilon, \mathscr{F}) \leqslant M_b(2, \mathscr{F}^{\gamma}).$$

The key of the proof of Theorem 4 is the following combinatorial lemma:

Lemma 5. *Let $\mathscr{X}$ be a set of cardinality m, and let $\mathscr{G}$ be a class of functions $\mathscr{X} \to \{1, \ldots, b\}$, where b is an even positive integer and $m \geqslant b/6$. Then*

$$M_b(2, \mathscr{G}) \leqslant 2(6m)^{\lceil \log_2 y \rceil},$$

where

$$y = \sum_{i=1}^{\Delta_b} \binom{m}{i} 2^i$$

and Δ_b *is the b-dimension of* $\mathcal{G}$.

Proof. We may assume that $b \geqslant 4$ since otherwise there are no two 2-separated functions in $\mathcal{G}$ and the statement is trivial. Let $A \subset \mathcal{X}$ and $s : A \to \{b/2, b/2+1\}$. We say that $\mathcal{G}$ b-shatters the pair (A,s) if it b-shatters A according to s. To any $k \geqslant 2$ and $m \geqslant 1$, define $t(k,m)$ as the largest integer t such that if $\mathcal{H}$ is any 2-separated class of functions with $|\mathcal{H}| = k$, then $\mathcal{H}$ b-shatters at least t distinct pairs (A,s). If no such $\mathcal{H}$ exists, then we say that $t(k,m) = \infty$. (Recall that $m = |\mathcal{X}|$.)

Clearly, the number of possible pairs (A,s) such that $|A| \leqslant d$ is at most $y = \sum_{i=1}^{d} \binom{m}{i} 2^i$. Thus, if $t(k,m) > y$ for some k, then $M_b(2,\mathcal{G}) < k$ whenever $\Delta_b \leqslant d$. Therefore, we need to show that $t\left(2(6m)^{\lfloor \log_2 y \rfloor}, m\right) > y$ for all $d \geqslant 1$, $m \geqslant 1$.

We see immediately that $t(2,m) = 1$ for all $m \geqslant 1$. Next, we show that

$$t(12km, m) \geqslant 2t(2k, m-1). \tag{13}$$

If there is no 2-separated class with size $12km$, then the left-hand side of (13) is ∞, and the inequality is trivially true. Thus, assume that there is a 2-separated class $\mathcal{H}$ with size $12km$. Split $\mathcal{H}$ into $6km$ pairs of functions. For each such pair (h_1, h_2),

$$\max_{x \in \mathcal{X}} |h_1(x) - h_2(x)| I_{\{u(h_1(x)) \neq u(h_2(x))\}} \geqslant 2$$

that is, for each such pair there exists an $x \in \mathcal{X}$ such that $|h_1(x) - h_2(x)| \geqslant 2$ and $u(h_1(x)) \neq u(h_2(x))$. Since $|\mathcal{X}| = m$, there exists an $x \in \mathcal{X}$ such that this property holds for at least $6k$ pairs. Order all these pairs such a way that $h_1(x) \geqslant h_2(x)$. For $j \in \{1, \ldots, b\}$, define

$$\tau(j) = \begin{cases} 1 & \text{if } j < b/2, \\ 2 & \text{if } j = b/2, \\ 3 & \text{if } j = b/2+1, \\ 4 & \text{if } j > b/2+1. \end{cases}$$

Since for all pairs $h_1(x) - h_2(x) \geqslant 2$ and $u(h_1(x)) \neq u(h_2(x))$, there are three possible values of the pair $(\tau(h_1(x)), \tau(h_2(x)))$ (namely (3,1), (4,1), and (4,2)). By the pigeonhole principle, there are at least $6k/3 = 2k$ pairs (h_1, h_2) for which the $(\tau(h_1(x)), \tau(h_2(x)))$ is the same. Then it follows that there are two subclasses $\mathcal{H}_1$, $\mathcal{H}_2 \subset \mathcal{H}$ and indices $i, j \in \{1,2,3,4\}$ with $|\mathcal{H}_1| = |\mathcal{H}_2| = 2k$ such that for each $h_1 \in \mathcal{H}_1$, $\tau(h_1(x)) = i$, for each $h_2 \in \mathcal{H}_2$, $\tau(h_2(x)) = j$, and $i \geqslant j+2$. Clearly, the members of $\mathcal{H}_1$ are 2-separated on $\mathcal{X} - \{x\}$, and the same is true for $\mathcal{H}_2$. Thus, according to the definition of $t(k,m)$, $\mathcal{H}_1$ and $\mathcal{H}_2$ both b-shatter $t(2k, m-1)$ pairs (A,s) with $A \subset \mathcal{X} - \{x\}$.

Clearly, $\mathscr{H}$ b-shatters every pair (A,s) which is b-shattered by either $\mathscr{H}_1$ or $\mathscr{H}_2$. Also, if a pair (A,s) is b-shattered by both $\mathscr{H}_1$ and $\mathscr{H}_2$, then it is easy to see that $\mathscr{H}$ b-shatters the pair $(A\cup\{x\},s')$, where $s'(z)=s(z)$ if $z\in A$, and

$$s'(x)=\begin{cases} b/2+1 & \text{if } j=2,\\ b/2 & \text{if } i=3,\\ \text{arbitrary} & \text{otherwise.}\end{cases}$$

Therefore, $\mathscr{H}$ b-shatters at least as many (A,s) pairs as the sum of the numbers of pairs shattered by $\mathscr{H}_1$ and $\mathscr{H}_2$, so (13) is proved.

Let now $n=2(6m)(6(m-1))\cdots(6(m-r+1))$, where $r\leqslant m$. Then by repeated application of (13), we obtain

$$t(n,m)\geqslant 2^r t(2,n-r)=2^r.$$

Since t is monotone in its first argument, for all $r\leqslant m$,

$$t(2(6m)^r,m)\geqslant 2^r.$$

Take $r=\lceil\log_2 y\rceil$. If $r\leqslant m$, then

$$t(2(6m)^{\lceil\log_2 y\rceil},m)\geqslant 2^{\lceil\log_2 y\rceil}>y,$$

as desired. If $r>m$, then $2(6m)^r>2(6m)^m>b^m$ by the condition $6m\geqslant b$. But b^m is the number of all functions from $\mathscr{X}$ to $\{1,\ldots,b\}$, so there is no 2-separated class larger than this, hence $t(2(6m)^{\lceil\log_2 y\rceil},m)=\infty>y$, establishing the lemma. □

Proof of Theorem 4. Let $b=4\lceil 1/\varepsilon+\frac{1}{2}\rceil$ and $\gamma=1/(b-1)$. Then it is easy to check that $\gamma\leqslant\varepsilon/4-$ so that the hypothesis of Lemma 4 is satisfied – and $\gamma/2\geqslant\varepsilon/10$. Then

$$\begin{aligned}
|\overline{\mathscr{F}_c}| &\leqslant M_b(2,\mathscr{F}^\gamma) \quad \text{(by Lemma 4)}\\
&\leqslant 2(6m)^{\lceil\log_2(\sum_{i=1}^{A_b}\binom{m}{i}2^i)\rceil} \quad \text{(by Lemma 5)}\\
&\leqslant 2(6m)^{\lceil\log_2(\sum_{i=1}^{d_{\gamma/2}}\binom{m}{i}2^i)\rceil} \quad \text{(by Lemma 3)}\\
&\leqslant 2(6m)^{\lceil\log_2(2em/d_{\gamma/2})^{d_{\gamma/2}}\rceil} \quad \left(\text{since } \binom{m}{i}2^i\leqslant\binom{2m}{i} \text{ and } \sum_{i-1}^{k}\binom{n}{i}\leqslant(ne/k)^k\right)\\
&\leqslant 2(6m)^{d_{\varepsilon/10}\log_2(2em/d_{\varepsilon/10})+1}
\end{aligned}$$

as desired. □

7. Concluding remarks

Several interesting questions have been left unanswered. Perhaps the most interesting one is the characterization of "learnability". We suspect that the finiteness of d_γ is

sufficient and necessary. More precisely, our conjecture is the following: given a class $\mathscr{F}$ of regression functions, a sample size of the order of (ignoring logarithmic factors)

$$\max\left(\frac{L^* d_\varepsilon}{\gamma^2}+\frac{L^*}{\gamma^2}\log\frac{1}{\delta},\frac{d_\varepsilon}{\gamma}+\frac{1}{\gamma}\log\frac{1}{\delta}\right)$$

is sufficient and necessary to guarantee the existence of a classifier g_n such that for all possible $\eta^* \in \mathscr{F}$ and for all distributions μ the excess probability of error $L(g_n)-L^*$ be smaller than γ with probability larger than $1-\delta$ for all δ and $\gamma > c\varepsilon$ for some constant c. We do not know the answer even if μ is allowed to be known. The bounds given here for classifiers $\hat{g}_n$ of Section 4 and $\bar{g}_n$ of Section 5 both fall short of this performance guarantee. In the case of known μ, $\hat{g}_n$ is a candidate to prove the sufficiency part. However, we have not been able to replace in Corollary 1 the minimum of V and P_ε by d_ε. The main technical difficulty comes from the fact that the "metric" which defines the covering does not satisfy the triangle inequality. In the case of unknown distributions, as we pointed it out, $\bar{g}_n$ fails in some cases when d_γ is finite. If our conjecture is true, a new classifier is to be found. Also, unlike in the completely distribution-free setup, the answers in the two cases (i.e., know and unknown μ) may be very different.

Acknowledgements

The insightful comments of the two reviewers helped us correct some key mistakes in the original draft and improve the presentation.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, Technical Report 143-95, DSI, University of Milan, Italy, 1993. An extended abstract appeared in the Proc. 1993 IEEE Symposium on the Foundations of Computer Science, IEEE Press, New York.

[2] M. Anthony, P.L. Bartlett, Function learning from interpolation, Technical report, Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia, 1994.

[3] M. Anthony, J. Shawe-Taylor, A result of Vapnik with applications, Discrete Appl. Math. 47 (1993) 207–217.

[4] P.L. Bartlett, The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, IEEE Trans. Inform. Theory (1997) to appear.

[5] P.L. Bartlett, S.R. Kulkarni, S.E. Posner, Covering numbers for real-valued function classes, Technical Report, Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia, 1996.

[6] P.L. Bartlett, P.M. Long, More theorems about scale-sensitive dimensions and learning, in: Proc. 8th Annual Conf. Computational Learning Theory, Association for Computing Machinery, New York, 1995, pp. 392–401.

[7] P.L. Bartlett, P.M. Long, R.C. Williamson, Fat-shattering and the learnability of real-valued functions, J. Comput. System Sci. 52 (1996) 434–452.

[8] G.M. Benedek, A. Itai, Learnability by fixed distributions, in: Computational Learning Theory: Proc. 1988 Workshop, San Mateo, CA, Morgan Kaufman, pp. 80–90, 1988.
[9] S.N. Bernstein, The Theory of Probabilities, Gastehizdat Publishing House, Moscow, 1946.
[10] A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, J. ACM 36 (1989) 929–965.
[11] K.L. Buescher, P.R. Kumar, Learning by canonical smooth estimation, Part I: Simultaneous estimation, IEEE Trans. Automat. Control 41 (1996) 545–556.
[12] K.L. Buescher, P.R. Kumar, Learning by canonical smooth estimation, Part II: learning and choice of model complexity, IEEE Trans. Automat. Control 41 (1996) 557–569.
[13] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.
[14] L. Devroye, G. Lugosi, Lower bounds in pattern recognition and learning, Pattern Recognition 28 (1995) 1011–1018.
[15] R.M. Dudley, S. Kulkarni, T. Richardson, O. Zeitouni, A metric entropy bound is not sufficient for learnability, IEEE Trans. Inform. Theory 40 (1994) 883–885.
[16] D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, Inform. Comput. 100 (1992) 78–150.
[17] M. Kearns, R.E. Shapire, Efficient distribution-free learning of probabilistic concepts, J. Comput. Sys. Sci. 48 (1994) 464–497.
[18] S.R. Kulkarni, Problems of computational and information complexity in machine vision and learning, Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1991.
[19] W.S. Lee, P. Bartlett, R.C. Williamson, Efficient agnostic learning of neural networks with bounded fan-in, IEEE Trans. Inform. Theory 42 (1996) 2118–2132.
[20] G. Lugosi, A. Nobel, Adaptive model selection using empirical complexities, 1996, submitted.
[21] G. Lugosi, M. Pintér, A data-dependent skeleton estimate for learning, in Proc. 9th Annual ACM Conf. Computational Learning Theory, Association for Computing Machinery, New York, 1996, pp. 51–56.
[22] A.B. Nobel, On uniform laws of averages, Ph.D. Thesis, Department of Statistics, Stanford University, Stanford, CA, 1992.
[23] D. Pollard, Convergence of Stochastic Processes, Springer, New York, 1984.
[24] D. Pollard, Rates of uniform almost sure convergence for empirical processes indexed by unbounded classes of functions, 1986, manuscript.
[25] N. Sauer, On the density of families of sets, J. Combinatorial Theory Ser. A 13 (1972) 145–147.
[26] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, M. Anthony, A framework for structural risk minimization, in Proc. 9th Annual Conf. Computational Learning Theory, Association of Computing Machinery, New York, 1996, p. 68–76.
[27] V.N. Vapnik, Estimation of Dependencies Based on Empirical Data, Springer, New York, 1982.
[28] V.N. Vapnik, A.Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (1971) 264–280.
[29] V.N. Vapnik, A.Ya. Chervonenkis, Theory of Pattern Recognition, Nauka, Moscow, 1974 (in Russian); German translation: Theorie der Zeichenerkennung, Akademie Verlag, Berlin, 1979.
[30] V.N. Vapnik, A.Ya. Chervonenkis, Necessary and sufficient conditions for the uniform convergence of means to their expectations, Theory Probab. Appl. 26 (1981) 821–832.